DOCUMENT RESUME

ED 175 875 TM 007 687

AUTHOR Moore, R. P.; And Others
TITLE Efficiency Study of NLS Base-Year Design. RTI-220-884-3.
INSTITUTION Research Triangle Inst., Durham, N.C. Statistics Research Div.
REPORT NO [illegible]T-220-884-3
PUB DATE [illegible] 74
NOTE 37p.

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Cost Effectiveness; *Data Analysis; *Efficiency; Followup Studies; Graduate Surveys; High School Graduates; *Longitudinal Studies; *National Surveys; Participant Characteristics; Postsecondary Education; Predictor Variables; *Research Design; *Sampling; Senior High Schools
IDENTIFIERS *National Longitudinal Study High School Class 1972; *Sample Size

ABSTRACT

An efficiency study was conducted of the base year design used for the National Longitudinal Study of the High School Class of 1972 (NLS). Finding the optimal design involved a search for the numbers of sample schools and students that would maximize the variance at a given cost. Twenty-one variables describing students' plans, attitudes, abilities, and self-concept were selected, and 17 sets of variance components were estimated for each variable. The efficiency of these 357 statistics for each of nine designs of approximately equal cost was computed, to yield an optimal design for all statistics. Optimum designs--those with the highest average cost efficiencies--were calculated for all statistics, national statistics, and domain statistics (statistics analyzed by sex, race, and father's education). The optimal design for national statistics involves 21 students sampled per school; optimal designs for domain statistics involve 24 white students sampled per school and 30 for blacks, other races, males, females, and father's education. The current NLS design, with 18 students per school, is not far from the optimum. Designs sampling 18 to 30 students per school all have high average efficiencies for both national and domain statistics. (CP)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

RESEARCH TRIANGLE INSTITUTE

ED175875

22U-884-3

EFFICIENCY STUDY OF NLS

BASE-YEAR DESIGN

by

R. P. Moore
B. V. Shah
R. E. Folsom

Statistics Research Division
Research Triangle Institute

November 1974

TM 007 687

RESEARCH TRIANGLE PARK, NORTH CAROLINA 27709

TABLE OF CONTENTS

		Page
LIST OF TABLES		iii
I.	INTRODUCTION	1
II.	OPTIMALITY CRITERION	1
III.	VARIABLES	3
IV.	VARIANCE COMPONENTS	6
V.	COST MODEL	12
VI.	RESULTS	20
	A. Efficiencies	20
	B. Sensitivity Analysis	26
	C. Variance Ratios	29
VII.	SUMMARY	29
REFERENCES		33

LIST OF TABLES

		Page
1.	Estimated Direct Charges Allocated to School and Student Costs, NLS Base Year Survey	16
2.	Estimated Man-months of Direct Labor Allocated to School and Student Costs, NLS Base Year Survey	16
3.	Calculation of Estimated Variable Cost Parameters C_1 and C_2, NLS Base Year Survey	17
4.	Specification of Nine Designs with Approximately Equal Cost	19
5.	Mean Efficiencies by Type of Statistic	21
6.	Standard Deviation of Efficiencies by Type of Statistic	22
7.	Mean Efficiencies for Domain Means and Differences between Domain and National Means	24
8.	Specification of Designs with Approximately Equal Cost for Alternative Cost Estimates for Sensitivity Analysis	25
9.	Cost Sensitivity of Mean Efficiencies for All Statistics	27
10.	Number of Respondents and Ratios of Variance Components, by Domains	28

I. INTRODUCTION

The sampling design for the National Longitudinal Study (NLS) of the Senior Class of 1972 may be viewed as a deeply stratified three-stage design. The three stages are (1) final stratum, (2) school, and (3) student. A total of eight stratification variables were used to define the 600 final strata. The sample design is described fully in a separate report [1]. For the purposes of this analysis, final stratum is regarded as a fixed stratification level, not subject to change. The problem of finding the optimal design involves a search for the numbers of sample schools and students which will result in minimum variance at a given cost. Since the cost is constrained, the independent parameter of the NLS design is the number of sample students per school.

Twenty-one selected base-year student variables and nine selected subpopulations were used for the analysis of sample efficiency. The analysis was carried out using the optimality criterion of Shah and Folsom [2]. Variance components were calculated using "VARCOMP", an RTI algorithm developed by Shah [3], using formulae written by Seeger [4] to estimate the components. The details of the formulae used in VARCOMP are shown in section IV. Cost components needed to estimate efficiencies were developed using available historical information and judgment. The results of the efficiency analysis are shown in section VI of this report.

II. OPTIMALITY CRITERION

If the objective of a sample survey is to estimate only one statistic, the usual procedure is to choose the design for which the variance of the

statistic is minimized for a given (fixed) survey cost. When many statistics are to be estimated from a single sample, the criterion for choosing the optimum design is not obvious. The possible criteria include (1) the design that is optimal for most statistics, (2) the design with minimum average variance at the given cost, and (3) the design with maximum average efficiency. The criterion of maximum average efficiency as described by Shah and Folsom [2] was used in the analysis of the NLS design.

Briefly stated, the problem involves defining a group of feasible designs with approximately equal cost, estimating the efficiency of each design for many statistics, and choosing as the "optimal" design the one with the highest average efficiency. A small variance of efficiencies is a desirable secondary criterion of the "optimal" design.

Suppose there are M designs D_k, ($k = 1, 2, \cdots, M$) and N statistics Y_ℓ, ($\ell = 1, 2, \cdots, N$). Let the estimated variance components for Y_ℓ be identified as $V(\ell,1)$ and $V(\ell,2)$ for the school and student components, respectively. Let C_1 be the marginal cost of adding another sample school, let C_2 be the marginal cost of adding another sample student, and assume a linear cost function. Then the efficiency of design k for statistic-ℓ is defined as

$$E(k,\ell) = \frac{\text{minimum variance for statistic-}\ell\text{ at the given cost}}{\text{variance of statistic-}\ell\text{ with design-k}}$$

where design-k has r_k sample schools and $r_k s_k$ sample students. The numerator and denominator of $E(k,\ell)$ may be written as

$$\text{numerator} = \frac{\left\{\sqrt{C_1 V(\ell,1)} + \sqrt{C_2 V(\ell,2)}\right\}^2}{C_1 r_k + C_2 r_k s_k},$$

and

$$\text{denominator} = \frac{V(\ell,1)}{r_k} + \frac{V(\ell,2)}{r_k s_k} \quad ,$$

respectively. Once the efficiencies were computed, the average efficiency, M_k, and the variance of the efficiencies, V_k, for each design were computed as

$$M_k = \frac{1}{N} \sum_{\ell=1}^{N} E(k,\ell) \quad ,$$

and

$$V_k = \frac{1}{N-1} \sum_{\ell=1}^{N} \left\{ E(k,\ell) - M_k \right\}^2 \quad .$$

The means and variances of efficiencies for each design are shown in section VI of this report. The variables selected for use in this analysis of design efficiency are discussed in the next section.

III. VARIABLES

A total of 21 base-year variables, 14 dichotomous variables and seven quantitative variables, were defined for the variance components project. The variables were selected on a judgmental basis as "important" in describing students' plans, attitudes, abilities, and self-concept. The variables selected also had to be applicable to all respondents and as a group had to reflect a wide range of the kinds of questions asked in the survey.

The 14 dichotomous variables selected are listed below. The numbers in parentheses are the values of the variable for the stated response.

1. Question 5: Which of the following best describes your grades so far in high school? Mostly B or above (1); below mostly B (0).

2. Question 12: Does your counselor usually have the information you feel you need? Yes (1); other responses (0).

3. Question 14A: How much have your parents influenced your plans for after high school? a great deal (1); somewhat or not at all (0).

4. Question 18D: Please rate your school on the aspect of job placement of graduates. poor (1); other responses (0).

5. Question 22A: Do you plan to use funds available from college or university scholarship or loan (athletic or academic) programs for further study beyond high school? Yes, I plan to use it (1); other responses (0).

6. Question 22B: Do you plan to use funds available from State or local scholarship or loan programs for further study beyond high school? Yes, I plan to use it (1); other responses (0).

7. Question 22F: Do you plan to use funds available from college work-study programs for further study beyond high school? Yes, I plan to use it (1); other responses (0).

8. Question 22R: Do you plan to use funds available from a regular bank loan for further study beyond high school? Yes, I plan to use it (1); other responses (0).

9. Question 23B: Here are some reasons others have given for not planning to get financial aid for study from sources outside the family. Which of these reasons apply to you? My parents or I will be able to pay for more education without getting any outside financial aid. Does not apply to me (1); does apply to me (0).

10. Question 24A: How important is making a lot of money to you in selecting a job or career? Very important (1); somewhat important or not important (0).

11. Question 24C: How important is the opportunity to be helpful to others or useful to society to you in selecting a job or career? Very important (1); somewhat important or not important (0).

12. Question 28: Whatever your plans, do you think you have the ability to complete college? Yes, definitely (1); other responses (0).

13. Question 31: What is the one thing that most likely will take the largest share of your time in the year after you leave high school? working full-time (1); other responses (0).

14. Question 31: What is the one thing that most likely will take the largest share of your time in the year after you leave high school? attending a four-year college or university full-time or part-time (1); other responses (0).

The seven quantitative variables used in the analysis of design efficiency were all composite variables formed by summing the responses to several questions or, in one case, by summing the scaled scores obtained on the six aptitude tests. The composite or index variables, expected to be used mainly for classifying students into homogeneous analysis subgroups, are described below:

1. Work orientation index - sum of the item responses to Question 20, parts A, C, and E.

2. Community orientation index - sum of the item responses to Question 20, parts F, G, and J.

3. Family orientation index - sum of the item responses to Question 20, parts B, H, and I, after reversing the scale values for item I.

4. Self-Concept index - sum of the item responses to Question 21, parts A, C, D, and H, after recoding "no opinion" as 3, "disagree" as 4, and "disagree strongly" as 5.

5. Control of environment index - sum of the item responses to Question 21, parts B, E, F, and G, after recoding "no opinion" as 3, "disagree" as 4, and "disagree strongly" as 5.

6. School problems index - sum of the item responses to Question 17, parts A, B, C, G, H, and I.

7. Aptitude index - sum of the test scores on the six aptitude tests in the test battery.

In addition to selecting the 21 variables, eight domains or subpopulations were chosen for analysis: males, females, whites, blacks, other races; and three levels of father's education: less than high school graduate, high school graduate, and some education after high school.

Seventeen statistics for each variable were estimated: a national mean, eight domain means, and eight differences between domain means and the national mean. Thus the total number of statistics computed was 357 or 17 x 21.

The next section describes the method used to compute estimates of the components of variance for each of the 357 statistics. In section VI, the means and variances of the 357 estimated efficiencies for each of the equal-cost design alternatives are compared.

IV. VARIANCE COMPONENTS

How best to define and estimate variance components for nonlinear statistics such as ratios and differences between ratios from complex multistage sample designs is still an open question. Some appropriate formulae using Taylor series and Jackknife linearizations have been proposed by Shah, Folsom, and Bayless [5]. Tayorized linearizations and Seeger [4] type variance components were selected for use in this analysis.

The NLS sample can be viewed as a nested sample with three levels. The first level is fixed and corresponds to the set of final strata used to stratify the schools in the sampling frame ($h = 1, 2, \cdots, H$). The two sampling levels nested within final strata are schools ($i = 1, 2, \cdots, r_h$) and students within schools ($j = 1, 2, \cdots, s_{hi}$). An expanded model with terms for each of the five major school stratification variables will be used in evaluating the stratification scheme. The magnitudes of the variance component estimates, and to a lesser extent the ratios of estimated components, are affected by the imbalance of the sample and by the order in which the fixed stratification variables are introduced in the model. Both the population and the sample are much more balanced with the three factor model than with the expanded model. Thus, the variance components estimated using the three-factor model are considered more appropriate for comparing the efficiencies of alternative sample designs.

The statistics of interest are ratios, means and proportions, for the variables described in section III.

Y_{hij} = the variate value for student (hij),

and

W_{hij} = the weight for student (hij),

then the estimated "National" mean is

$$\hat{\bar{Y}}_N = \sum_{h=1}^{H} \sum_{i=1}^{r_h} \sum_{j=1}^{s_{hi}} W_{hij} Y_{hij} \Big/ W_{+++} \quad ,$$

where W_{+++} is the sum of the weights over (hij)*. For variance estimation purposes, we used the linearized values:

---

*In general we will use (+) to indicate a sum over a particular subscript while (·) will denote the average over that subscript.

$$TN_{hij} = W_{hij}\left(Y_{hij} - \hat{\bar{Y}}_N\right)\Big/W_{+++} \quad .$$

For specific subpopulations or domains, such as blacks, we used indicator variables such as

$$X_{hij} = \begin{cases} 1 \text{ if student (hij) was black, or} \\ 0 \text{ otherwise;} \end{cases}$$

and estimated the domain mean as

$$\hat{\bar{Y}}_B = \sum_{hij} X_{hij}\, W_{hij}\, Y_{hij} \Big/ \sum_{hij} X_{hij}\, W_{hij} \quad .$$

The corresponding linearization for the black mean is

$$TB_{hij} = X_{hij}\, W_{hij}\left(Y_{hij} - \hat{\bar{Y}}_B\right)\Big/XW_{+++} \quad ,$$

where XW_{+++} is the denominator of $\hat{\bar{Y}}_B$, the mean of variable Y for the blacks.

The other statistics considered were differences between domain and national means of the form

$$\Delta\,\hat{\bar{Y}}_B = \left(\hat{\bar{Y}}_B - \hat{\bar{Y}}_N\right) \quad .$$

The corresponding linearization is

$$T\Delta B_{hij} = \left(TB_{hij} - TN_{hij}\right) \quad .$$

To facilitate the specification of our variance components model, we define

$$Z\Theta_{hij} = \hat{\Theta} + Hr_h s_{hi}\, T\Theta_{hij} \quad ,$$

where $\hat{\Theta}$ is either $\hat{\bar{Y}}_N$, $\hat{\bar{Y}}_B$, or $\Delta\hat{\bar{Y}}_B$ and $T\Theta_{hij}$ is the corresponding TN_{hij}, TB_{hij}, or $T\Delta B_{hij}$ value. The coefficients of $T\Theta$ make it possible to write the estimator $\hat{\Theta}$ as the simple average of the linearized Z variables; that is

$$\hat{\theta} = Z\theta_{...} = \frac{1}{H} \sum_{h} \frac{1}{r_h} \sum_{i} \frac{1}{s_i} \sum_{j} Z\theta_{hi} \quad .$$

Our variance components are defined in terms of the following linear model for the Z variables:

$$Z_{hij} = \mu + a_h + b_{hi} + c_{hij}$$

The variance components are defined as follows:

$$\sigma_A^2 = \frac{1}{H-1} \sum_{h=1}^{H} \left(a_h - a_.\right)^2 \quad ,$$

$$\sigma_B^2 = \frac{1}{H} \sum_{h=1}^{H} \frac{1}{R_h - 1} \sum_{i=1}^{R_h} \left(b_{hi} - h_{h.}\right)^2 \quad ,$$

and

$$\sigma_C^2 = \frac{1}{H} \sum_{h=1}^{H} \frac{1}{R_h} \sum_{i=1}^{R_h} \frac{1}{S_{hi} - 1} \sum_{j=1}^{S_{hi}} \left(c_{hij} - c_{hi.}\right)^2 \quad .$$

In the definition of these variance components, capital R's and S's have been used to indicate the population sizes, number of schools per final stratum and number of seniors per school, for the two sampled levels. Since the first stage is fixed and the last two are random, it can be shown that the variance of $\hat{\theta}$, for a design-k with r_k schools and s_k students per school can be written as

$$V(\hat{\theta}) = \frac{\sigma_B^2}{r_k} + \frac{\sigma_C^2}{r_k s_k} \quad .$$

The mean squares proposed by Seeger [4] for these designs are of the form:

$$MSC^1 = \frac{1}{H} \sum_{h=1}^{H} \frac{1}{r_h} \sum_{i=1}^{r_h} ms(hi) \quad ,$$

where ms(hi) is the between student within school mean square among the $Z\Theta_{hij}$ variates; namely

$$ms(hi) = \sum_{j=1}^{s_{hi}} \left(Z\Theta_{hij} - Z\Theta_{hi.}\right)^2 \Big/ \left(r_{hij} - 1\right).$$

A more convenient computational form of MSC^1 is

$$MSC^1 = H \sum_{h=1}^{H} r_h \sum_{i=1}^{r_h} s_{hi}^2 \, MS(hi) \quad ,$$

with

$$MS(hi) = \left(\sum_{j=1}^{s_{hi}} T\Theta_{hij}^2 - T\Theta_{hi+}^2 \Big/ s_{hi} \right) \Big/ \left(s_{hi} - 1 \right) \quad .$$

To estimate the school component of variance, we will also need

$$MSC^2 = H \sum_{h=1}^{H} r_h \sum_{i=1}^{r_h} s_{hi} \, MS(hi) \quad .$$

Continuing on, we will need

$$MSB^1 = H \sum_{h=1}^{H} r_h^2 \, MS(h) \quad ,$$

and

$$MSB^2 = H \sum_{h=1}^{H} r_h \, MS(h) \quad ,$$

where

$$MS(h) = \left(\sum_{i=1}^{r_h} T\Theta_{hi+}^2 - T\Theta_{h++}^2 \Big/ r_h \right) \Big/ \left(r_h - 1 \right) \quad .$$

For the fixed or stratification level we use

$$MSA = H^2\left(\sum_{h=1}^{H} T\theta_{h++}^2 - T\theta_{+++}^2 \Big/ H\right)\Big/\left(H - 1\right) \quad .$$

After calculating the set of five mean squares, the variance components are estimated by

$$\hat{\sigma}_A^2 = MSA - MSB^2 ,$$

$$\hat{\sigma}_B^2 = MSB^1 - MSC^2 ,$$

and

$$\hat{\sigma}_C^2 = MSC^1 .$$

The variance component definitions presented in this section are specified in terms of simple random selections at the school and student stages. Seeger [4] also points out in this case that the estimators above are biased unless the finite population corrections at each sampling stage are negligible, or the number of units in the population is considerably greater than the sample size at that level. The NLS design incorporates PPS with replacement selection at the school stage. It happens that the same mean square calculations with properly weighted variables also provide sample estimators for variance components based on unequal probability selections. Folsom, Bayless, and Shah [5] have shown for such selections that the true expectation of Seeger's component estimators represent composites of the sampling variation at the level indicated minus a finite population correction from the next lowest level in the hierarchy. Fortunately, these composite components, say Σ_B^2 and Σ_C^2, can be used to model the true variance of a statistic $\hat{\theta}$ from an NLS-like design. The model takes the form

$$\text{Var}\ \{\hat{\theta}\} = \Sigma_B^2 / r + \Sigma_C^2 / rs - \Sigma_0^2$$

$$= V_{\hat{\theta}}^2\ (r,s) - \Sigma_0^2$$

where Σ_0^2 is the finite population correction from the school selection stage. Our variance components model will estimate the variable portion of Var $\{\hat{\theta}\}$, the portion $V_{\hat{\theta}}^2\ (r,s)$ which depends on r and s. Using an estimate of

$$E_{\hat{\theta}}\ (k) = \left[V_{\hat{\theta}}^2\ (\min) \Big/ V_{\hat{\theta}}^2\ (r_k, s_k) \right] ,$$

instead of the true efficiency

$$E_{\theta}\ (i) = \left[V_{\hat{\theta}}^2\ (\min) - \Sigma_0^2 \right] \Big/ \left[V_{\hat{\theta}}^2\ (r_k, s_k) - \Sigma_0^2 \right] ,$$

results in an overstatement of the efficiency of each design-i. In spite of this overstatement, the relative ranking of designs will not be altered.

V. COST MODEL

The cost model for the variance components analysis was assumed to be linear of the form

$$C = C_0 + n_1 C_1 + n_1 n_2 C_2 ,$$

where

- C = total cost of the base-year survey;
- C_0 = portion of the cost which is fixed and does not depend on the sample design;
- C_1 = marginal cost of adding an additional school to the sample;
- C_2 = marginal cost of adding another student to the sample in a school;
- n_1 = number of sample schools, and
- n_2 = number of sample students per school.

The parameter C_0 is assumed to include costs of central staff administrative and some technical work; costs of space and equipment; presurvey costs for constructing questionnaires, preparing instruction manuals, discussing survey objectives, designing the sample, and selecting the sample of schools; and post-survey costs for analyzing the results, preparing reports, and distributing the survey reports to the educational community.

Among the costs normally included in the parameter C_1 are the costs of printing school and counselor questionnaires, selecting the sample of students, shipping materials to the schools, telephone calls to the schools, and costs of receiving data from the schools.

Costs classified as variable costs associated with an additional sample student within a school include the cost of printing student questionnaires, shipping costs, reviewing and editing the completed questionnaires, and converting the collected information to a machine-readable form.

In order to perform the variance components analysis and compare a number of cost-equivalent design alternatives, it was necessary to obtain estimates of the magnitudes of C_1, and C_2, the parameters involved with the variable portion of the survey cost.

The 13 major tasks involved in carrying out the 1972 NLS base-year project were as follows:

task 1 - relating the study to the educational community;

task 2 - sample selection and weighting;

task 3 - securing cooperation and participation in the study;

task 4 - design and procurement of study materials;

task 5 - distribution of study materials and collection of data;

task 6 - receipt control and followup actions;

task 7 - data editing and followup actions;

task 8 - data base and software support;

task 9 - data analysis;

task 10 - quality evaluation of field administration;

task 11 - reliability and validity of key student questionnaire items;

task 12 - reporting study results to State education agencies, and

task 13 - reports to the Office of Education.

In order to estimate the variable cost portion of the survey cost, it was necessary to decide which tasks involved variable costs and which involved fixed costs. Tasks 1, 8, 9, 10, 11, 12, and 13 were considered to involve only fixed costs within the range of alternative sample designs appropriate for the NLS project. Except for the printing costs, task 4 was also considered a fixed cost task. These judgments were supported by RTI's initial estimates of the man-months of labor required for the 900-school and 1200-school samples for each of these base-year tasks (table 1 of RTI Proposal Number 22-72-03 dated December 27, 1971). That is, the estimated level of effort for professional and support personnel was the same, or nearly so, for each of the tasks considered as fixed. The costs necessary for project management were also considered as fixed. Thus, the variable portion of the total cost was assumed to be associated with some of the costs necessary to perform tasks 2, 3, 4, 5, 6, and 7.

Cost data obtained from the Office of Education for a number of direct cost items were used to estimate the direct charges portion of variable costs.

The RTI base-year proposal estimated printing costs indicated that 20 percent of the printing costs were for school and counselor forms while 80 percent was for student questionnaires and answer sheets. Printing and shipping costs were allocated in this ratio. However, the school and counselor questionnaires accounted for only five percent of the forms and two percent of the total pages. Thus, data processing costs for editing and keytaping were allocated to school- and student-related costs by assigning 10 percent to school and 90 percent to student costs. The data processing costs for selecting the student samples and the telephone and communications costs were considered as school-related costs only (see table 1).

The portion of the direct labor costs for tasks 2 through 7 which was considered to be variable rather than fixed was estimated from the original RTI proposal and based on experience. The variable portion of the estimated labor requirements are shown in table 2, by task. Professional and support labor requirements were estimated separately. No estimate is shown for task 4 since the labor was considered as a fixed cost item. Direct labor for tasks 2, 3, and 5 were considered school-related costs while that for tasks 6 and 7 was allocated to school and student cost components, 10 percent as school costs and 90 percent as student costs.

The cost parameters, C_1 and C_2, relating to the school- and student-related variable costs were estimated based on the data assembled in tables 1 and 2. Table 3 shows how the estimates were calculated. Direct labor related to variable costs was converted to an estimate in dollars using approximate cost factors of $1,500 per professional man-month and $600 per support man-month.

Table 1 - Estimated Direct Charges Allocated to School and Student Costs, NLS Base Year Survey

Item	School cost (C_1) dollars	Student cost (C_2) dollars
Printing	5,280	21,120
Postage and shipping	3,700	14,800
Telephone and communications	14,000	-
School honorariums	30,000	30,000
Data processing:		
Selecting student samples	1,300	-
Editing and keytaping	3,570	32,130
Total Direct Charges	57,850	98,050

Table 2 - Estimated Man-months of Direct Labor Allocated to School and Student Costs, NLS Base Year Survey

Item	School cost (C_1)		Student cost (C_2)	
	Professional man-months	Support man-months	Professional man-months	Support man-months
Task 2	2.0	7.0	-	-
Task 3	5.0	10.0	-	-
Task 4	-	-	-	-
Task 5	7.0	9.0	-	-
Task 6	.7	3.5	6.3	31.5
Task 7	.8	4.5	7.2	40.5
Total	15.5	34.0	13.5	72.0

Table 3 - Calculation of Estimated Variable Cost Parameters C_1 and C_2, NLS
Base Year Survey

Item	School cost (C_1) dollars	Student cost (C_2) dollars
Direct labor		
Professional	23,250	20,250
Support	20,400	43,200
Total	43,650	63,450
Indirect technical expense[1/]	32,738	47,588
Direct charges	57,850	98,050
Overall management and service expense[2/]	14,766	23,000
Total estimated costs	149,004	232,088
Fixed fee[3/]	7,450	11,604
Total estimated cost plus fixed fee	156,454	243,692
Estimated cost parameters	130.38	11.28

1/ Computed as 75 percent of direct labor.

2/ Computed as 11 percent of direct labor, indirect technical expense, and direct charges.

3/ Calculated as five percent of total estimated costs.

Other direct costs as estimated in table 1 were used. Overhead or indirect costs were included as indirect technical expense (75 percent of direct labor) and overall management and service expense (11 percent of the total estimated direct and indirect cost). A fixed fee of five percent was also included to arrive at estimated total variable costs of $156,454 for school costs and $243,692 for student costs (table 3). Dividing the school cost estimate by 1,200 and the student cost estimate by 21,600 results in the estimated cost parameters, C_1 = $130.38 and C_1 = $11.28 shown in table 3.

RTI has been informed that a school honorarium may not be paid in subsequent NLS base-year projects. The estimated school cost plus fixed fee in table 3 would be reduced by 34,965 or 30,000 x 1.11 x 1.05, which would reduce the estimate for C_1 to $101.24. The estimated student cost plus fixed fee would also be reduced by 34,965 which would reduce the estimate for C_2 to $9.66.

Since the decision regarding the honorarium had not been made, average values of C_1 = $115 and C_2 = $10.50 were used to develop the nine designs shown in table 4. The designs shown were obtained by varying n_2, the number of students per school, and solving for the n_1, the number of sample schools, which would approximately equalize the total variable cost, $C - C_0$.

Table 4. Specification of Nine Designs with Approximately Equal Cost*

Design	Number of schools (n_1)	Number of students per school (n_2)	Total no. of students (n_1 n_2)	Total variable cost ($C - C_0$)
1	2,049	6	12,294	364,722
2	1,741	9	15,669	364,740
3	1,513	12	18,156	364,633
4	1,338	15	20,070	364,605
5**	1,200	18	21,600	364,800
6	1,087	21	22,827	364,688
7	994	24	23,856	364,798
8	915	27	24,705	364,628
9	848	30	25,440	364,640

*With C_1 = \$115 and C_2 = \$10.50

**Design used in the base year.

VI. RESULTS

A. Efficiencies

A total of 21 variables were selected for the study, as described in section III. For each variable, 17 sets of variance components were estimated: one set associated with the national mean, eight sets associated with domain means, and eight sets associated with differences between the domain means and the national mean. (Domains or subpopulations used were males, females, whites, blacks, other races; and three levels of father's education: less than high school graduate, high school graduate, and some education beyond high school.) The 357 sets of variance components were estimated using the methods described in section IV.

The estimated efficiency of statistic-ℓ for design-k ($\ell = 1, 2, \ldots, 357$; $k = 1, 2, \ldots, 9$) was computed using the formula for $E(k,\ell)$ given in section II, except that zero was substituted for the variance components with negative estimates. The means and variances of the efficiencies for each of the nine equal-cost designs were then computed as described in section II.

The mean efficiencies for each design are shown in table 5, for all statisti and by type of statistic. Design 6 is the "optimal" design for national level statistics; designs 8 and 9 are "optimal" for domain means; and design 9 is the "optimal" design for estimating differences between domain and national means. Design 9 also is "optimal" for all statistics if one is willing to use this optimality criterion and assign equal importance to each of the 357 statistics used in the study. Alternatively, one might assign equal importance to national means and domain means and no importance to differences of domain means and national means. In this situation, one might select design 7 as the "optimal"

Table 5. Mean Efficiencies by Type of Statistic

Design	Type of Statistic			
	National mean	Domain mean	Difference of domain and national mean	All
1	.721	.613	.547	.588
2	.850	.746	.676	.720
3	.919	.830	.762	.803
4	.953	.883	.820	.858
5	.968	.916	.861	.893
6	.970	.935	.889	.915
7	.964	.946	.908	.929
8	.953	.951	.920	.937
9	.939	.951	.928	.940
Number of efficiencies	21	168	168	357

Table 6. Standard Deviations of Efficiencies by Type of Statistic

Design	Type of Statistic			
	National mean	Domain mean	Difference of domain and national mean	All
1	.113	.121	.122	.129
2	.098	.128	.135	.138
3	.076	.124	.137	.135
4	.055	.116	.133	.128
5	.040	.106	.127	.118
6	.036	.096	.120	.109
7	.043	.088	.112	.100
8	.054	.081	.105	.093
9	.066	.076	.098	.087
Number of efficiencies	21	168	168	357

design by looking at the averages of the mean efficiencies in the first two columns of table 5.

Actually, several of the nine designs considered have high mean efficiencies. For example, designs 5, 6, 7, 8, and 9 each have mean efficiencies higher than .86 for each of the three types of statistics shown in table 5. The data of table 5 indicate, based on the cost modeling in section V and the optimality criterion used, that the "optimal" design for the NLS base-year survey is one of the designs 6, 7, 8, or 9. The specific choice depends on the relative importance given to the three types of statistics shown in table 5. The current design, number 5, is fairly near the optimum.

Table 6 shows the standard deviation of the efficiencies for the nine equal-cost designs, again by type of statistic. Comparison of tables 5 and 6 shows that, for the designs considered, a higher mean efficiency is associated with a lower standard deviation of efficiency. Thus, the secondary criterion of a small standard deviation of efficiencies does not seem to be in conflict with the primary criterion of high mean efficiency.

The mean efficiencies for each design for each of the eight domains used in the study are shown in table 7. Each of the means shown is based on 42 efficiencies, 21 domain means and 21 differences of the domain means and the corresponding national means. Design 9 is the "optimal" design for all subpopulations except whites. Designs 5, 6, 7, and 8 all have mean efficiencies greater than .85 for each of the eight subpopulations. Again, one needs some way of assigning importance to national means and domain means. Perhaps it is sufficient to say that designs 5, 6, 7, 8, and 9 are all highly efficient for national statistics and all subpopulation statistics considered in the study.

Table 7. Mean Efficiencies for Domain Means and Differences between Domain and National Means

Design	Race			Sex		Father's education		
	Whites	Blacks	Other	Males	Females	Less than high school	High school	Beyond high school
1	.663	.596	.577	.571	.593	.536	.565	.537
2	.796	.727	.710	.702	.725	.666	.697	.668
3	.873	.809	.795	.788	.809	.754	.784	.756
4	.918	.861	.851	.844	.863	.815	.842	.818
5	.943	.894	.888	.882	.899	.857	.882	.861
6	.954	.915	.912	.907	.921	.887	.908	.892
7	.957	.926	.927	.923	.935	.909	.925	.914
8	.955	.931	.935	.932	.942	.923	.936	.930
9	.948	.932	.939	.937	.945	.933	.942	.940
Number of efficiencies	42	42	42	42	42	42	42	42

Table 8. Specification of Designs with Approximately Equal Cost for Alternative Cost Estimates for Sensitivity Analysis

Design	C_1 = \$92, C_2 = \$12.60		C_1 = \$115, C_2 = \$10.50		C_1 = \$138, C_2 = \$8.40	
	Number of schools	Number of students per school	Number of schools	Number of students per school	Number of schools	Number of students per school
1	2,282	6	2,049	6	1,842	6
2	1,862	9	1,741	9	1,624	9
3	1,573	12	1,513	12	1,453	12
4	1,361	15	1,338	15	1,314	15
5	1,200	18	1,200	18	1,200	18
6	1,072	21	1,087	21	1,103	21
7	969	24	994	24	1,021	24
8	885	27	915	27	951	27
9	813	30	848	30	889	30

B. Sensitivity Analysis

The cost modeling described in section IV was done in a careful manner, considering all available cost data and the opinions of various knowledgeable persons. Separation of the variable costs into school-related costs and student-related costs, however, involved subjective judgments and the allocation could be questioned. The authors feel that the estimated cost components are probably accurate within 20 percent, plus or minus. An analysis was carried out in order to test the sensitivity of the "optimal" design to possible errors of 20 percent in the estimated cost parameters.

If one of the cost parameters is increased by 20 percent and the other decreased by 20 percent, the ratio of the two cost parameters will be changed by 1.5 or 0.67. The efficiencies were recalculated for two new sets of nine equal-cost designs: one set derived using C_1 = \$92.00 and C_2 = \$12.60; the second using C_1 = \$138.00 and C_2 = \$8.40. Table 8 shows the two sets of designs obtained with the above cost parameters. The set of designs developed in table 4 is repeated for comparison. The two new sets of designs were derived by keeping the number of students per school the same and solving for the number of schools necessary to hold the total variable cost approximately equal to that of the base-year design ($1{,}200\ C_1 + 21{,}600\ C_2$).

Next the efficiencies were recalculated using the new designs and the variance components estimated previously. Table 9 shows the mean efficiencies for each design with the three sets of estimated cost components. The "optimal" design for a cost ratio of 7.3 appears to be design 7 or design 8, while design 9 is "optimal" for cost ratios of 11.0 or 16.4. Thus, it appears that the "optimal" design is not very sensitive to errors of 20 percent or less in the

Table 9. Cost Sensitivity of Mean Efficiencies for All Statistics

Design	Cost parameters		
	C_1	92	115	138
	C_2	12.60	10.50	8.40
	C_1/C_2	7.3	11.0	16.4
1		.680	.588	.498
2		.800	.720	.632
3		.868	.803	.725
4		.908	.858	.791
5		.930	.893	.838
6		.942	.915	.872
7		.946	.929	.895
8		.946	.937	.911
9		.942	.940	.922
Number of efficiencies		357	357	357

Table 10. Number of Respondents and Ratios of Variance Components, by Domains

Domain	Average number of domain respondents	Percent of within stratum variation	
		School	Student
All students	16,024	3.1	96.9
Whites	12,215	2.2	97.8
Females	8,021	1.4	98.6
Males	7,877	1.2	98.8
Father - high school graduate	6,674	1.0	99.0
Father - less than high school	4,851	0.8	99.2
Father - beyond high school	2,545	0.8	99.2
Blacks	1,969	1.5	98.5
Other races	1,528	1.2	98.8

Note: Variance ratios for all students are averages for 21 national means; variance ratios for all domains are averages for 42 statistics-- 21 domain means and 21 differences between domain and national means.

C. Variance Ratios

Variance components were estimated for 357 statistics, as described earlier in this report. Averages of the component estimates would not be meaningful since the variables are measured on different scales involving different units. Ratios of the two random components of variance may be of interest, however, in the design of future studies.

Ratios of the school component to the sum of the school and student components were computed for each of the 357 statistics. The complement, student component divided by the sum of the school and student components, was also calculated for each statistic. Average values of these two ratios, expressed as percentages, are shown in table 10 by type of statistic. The student component percentage for national statistics is 96.9 percent, on the average. There appears to be a tendency for the school component to be a larger portion of the variance for the larger domains. Also, the school components is slightly larger for domains defined by race than for those defined by father's education.

VII. SUMMARY

The steps involved in the analysis of the efficiency of the base-year sample design are listed here as an aid in summarizing the study:

(a) Step 1 - A cost model was developed which expresses the total variable costs of the survey as a linear function of the numbers of schools and students sampled.

(b) Step 2 - A set of nine feasible sample designs with approximately equal total variable costs were derived using the cost parameters estimated in step 1.

(c) Step 3 - Twenty-one base-year variables and eight domains defined by race, sex, and father's education were chosen for the analysis.

(d) Step 4 - Variance components were calculated using Seeger's formulae and Taylorized linearizations for each of 357 statistics defined by the variables and domains chosen for the study.

(e) Step 5 - The efficiency of each statistic ($\ell = 1, 2, \ldots, 357$) for each of the sample designs ($k = 1, 2, \ldots, 9$) was computed as the minimum variance of the statistic at the given cost divided by the variance of the statistic with design-k.

(f) Step 6 - The average efficiencies of each design were calculated for all statistics, national statistics, and domain statistics.

(g) Step 7 - The "optimum" designs for all statistics, national statistics, and domain statistics were chosen as the designs with the highest average efficiencies.

The assumptions made in the analysis are briefly listed below to aid in interpreting the results of the study:

(a) The model used to express the total variable cost of the survey was assumed to be a linear function of the numbers of schools and students sampled.

(b) The cost model included only the costs of conducting the base-year survey.

(c) The first order Taylor series linearization provided a reasonable approximation to the variance of ratio estimates and differences between ratios.

(d) The first-stage sampling units, schools, were treated as if they were selected with varying probabilities with replacement (or without replacement from a large population) for variance components purposes.

(e) Negative variance component estimates were set to zero, assuming that the associated population values were negligible.

Assumption a implies that the results might not be valid over a wide range of values for r_k and s_k, but the model should be adequate for determining cost-equivalent designs in the neighborhood of the current design.

Assumption b is considered the most serious of all the assumptions, since it might be more appropriate to optimize the sample design for an entire NLS project including several followup surveys. An analysis involving this type of cost model and statistics defined as changes from base-year to first followup is now underway.

Assumptions c, d, and e are not considered serious. The component variance model is positively biased due to ignoring the finite population correction term at the school selection stage and due to the fact that negative school component estimates were set equal to zeros. The lack of balance in the sample and the population across final strata also inflates the component estimates somewhat. Thus, the relative efficiency measures are inflated but the efficiency ranking of the competing designs is not altered.

The following points summarize the principal results of the NLS base-year efficiency study:

(a) The "optimal" design for national statistics involves 21 sample students per school.

(b) The "optimal" designs for domain statistics involve 24 students per school for white students and approximately 30 per school for the other domains included in the study.

(c) The current NLS design with 18 students per school is not far from the "optimal" design.

(d) Designs with from 18 to 30 sample students per school all have high average efficiencies for both national and domain statistics.

(e) The criteria of maximum mean efficiency and minimum standard deviation of efficiencies appear to lead to the same "optimum" design choice.

(f) The "optimal" designs are not very sensitive to errors in the estimated cost component ratio, C_1/C_2, of from 0.67 to 1.50.

(g) The ratios of the school component estimate to the sum of school and student component estimates averaged approximately 97 percent for national statistics and from 98 to 99 percent for domain statistics.

REFERENCES

[1] "Sample Design for the Selection of a Sample of Schools with Twelfth-Graders for a Longitudinal Study," unpublished manuscript, Westat, Inc., Rockville, Md., June 1972.

[2] Shah, B. V. and R. E. Folsom, "Optimum Cluster Sizes in Multi-Purpose Sample Surveys," Proceedings of the Social Statistics Section, American Statistical Association, 1973.

[3] Shah, B. V. and C. N. Dillard, VARCOMP Program for Computing Variance Components, Working Paper No. 30, RTI Project 255U-796, 1973.

[4] Seeger, Paul, "A Method of Estimating Variance Components in Unbalanced Designs," Technometrics, 12, 1970, pp. 207-218.

[5] Folsom, R. E., D. L. Bayless, and B. V. Shah, "Jack-Knifing for Variance Components in Complex Sample Survey Designs," Proceedings of the Social Statistics Section, American Statistical Association, 1971.